DOCUMENT RESUME

ED 211 700 CE 030 943

AUTHOR Berube, Jean E.; Mark, Jorie Lester, Ed.
TITLE On Adult Learning. Measures of Effectiveness for Validation of an Experimental Design.
INSTITUTION Office of Vocational and Adult Education (ED), Washington, D.C.
PUB DATE Oct 81
NOTE 24p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adult Education; *Demonstration Programs; Educational Improvement; *Evaluation Criteria; *Evaluation Methods; *Program Effectiveness; Reliability; Statistical Significance; Validity
IDENTIFIERS Impact; *Joint Dissemination Review Panel; Objectivity; Replication

ABSTRACT

Evaluation design is discussed in terms of conditions that an adult education intervention (product, practice) must meet to get Joint Dissemination and Review Panel (JDRP) approval. (Effectiveness, the sole criterion for JDRP approval, must be established by evaluation data adequate to tie the project and desired impact together in a cause-and-effect relationship.) Four conditions examined by the JDRP are considered: (1) the evidence must be valid and reliable, (2) the effect must be of sufficient magnitude and have educational importance, (3) it should be possible to reproduce both the intervention and its effects at other sites, and (4) project data must be believable and interpretable. Discussion of statistical significance are size effect, importance of the educational area, and cost of the intervention. Considerations for replicability include setting, staff, participants, and components. Topics under the final condition of believability and interpretability include consistency of factual data in narrative and tables, completeness of data, and objectivity maintained in gathering data. An evaluation design checklist is appended. (YLB)

ON ADULT LEARNING

UNITED STATES DEPARTMENT OF EDUCATION
OFFICE OF VOCATIONAL & ADULT EDUCATION
WASHINGTON, D.C. 20202 - 3586

MEASURES OF EFFECTIVENESS FOR VALIDATION OF AN EXPERIMENTAL DESIGN

by

Jean E. Berube
Education Program Specialist

Edited by

Dr. Jorie Lester Mark
Chief
Development and Dissemination Branch
Adult Learning and Community Education, OVAE
U.S. Department of Education

October, 1981

TABLE OF CONTENTS

INTRODUCTION

The problem is that of proving the effectiveness of adult education interventions (i.e., products, or practices) to the Joint Dissemination and Review Panel (J.D.R.P.) of the U.S. Department of Education.

In order to be endorsed by J.D.R.P. for the Department of Education, educational interventions must be shown to have positive impacts on their recipients. Those positive impacts may be educational, attitudinal or behavorial in nature.

Effectiveness is the sole criterion for approval by J.D.R.P. In order to establish effectiveness there must be evaluation data adequate to tie the project and the desired impact together in a cause-and-effect relationship. To get J.D.R.P. approval an intervention must meet several conditions:

I. The evidence must be valid and reliable: Statistical Significance
II. The effect must be of sufficient magnitude and have educational importance: Educational Significance
III. It should be possible to reproduce both the interventions and its effects at other sites: Replicability
IV. Project data must be believable and interpretable: Believability and Interpretability

I. STATISTICAL SIGNIFICANCE

The main idea in evaluating an exemplary program is to measure the intended positive effect which was achieved because of the intervention and which was not compromised because of side effects. The measure(s) used must be

statistically valid and reliable in order to establish statistical significance.

A. VALIDITY

1. Logical Measurement

A valid measure is one that relates specifically to a certain kind of change. This change can be educational, attitudinal or behavorial and each kind of change requires a different kind of test. The measure selected must bear a logical relationship to the specific behavior being examined. If measures are unrelated to the behaviors that a program seeks to change, it is impossible to draw accurate conclusions about program effectiveness. For example, a program developed to improve reading skills cannot use in its evaluation measures of self-concept, mathematics skills or attitudes towards education because these measures do not have a direct relationship to reading skills.

2. Uncompromised by side effects

The effect or effects must also be shown to be uncompromised by side effects. Some side effects which must be considered and rejected are:

a. Side effects from the experiment.

b. Other simultaneous innovations.

c. Changes in population.

d. Differences in growth factors.

e. Maturation or simple passage of time.

All of these possibilities should be considered in interpreting evaluation results and, as far as possible, reasons should be presented why whatever gains were observed should be attributed to the treatment and not to such other influences as the side effects just listed. It is possible to make provisions in the design of the evaluation to negate many of these alternative explanations. For this reason it is highly desirable to obtain the services of a professional evaluator at the very beginning of a project so that a proper evaluation design may be created.

If a control or comparison group is used, every effort must be made to ensure that its members are as similar as possible to those in the treatment group. Systematic differences between groups in other factors such as urban or rural environment, race, socioeconomic status, or sex may be related to school performance. If these educationally relevant factors are not similar for both groups, it is difficult to determine whether the observed differences resulted from the intervention, or from differences in these other factors.

An Example of Convincing Evidence:

Performance in other areas can serve as one indication of change or consistency in the student population. For example, a new reading program in the small town of Andover, Massachusetts, had apparently produced a significant improvement in the performance of the students on district-wide standardized reading tests. The question was whether the effect might have been the result of an influx of higher achieving students into the district. The evaluators stated that there had been no perceptible change in the composition of the population over the previous three

years. To support this statement, they pointed out that the new program emphasized reading comprehension, and there had been large gains on the comprehension, subtests. However, performance on word attack skills, emphasized in both the old and new programs, remained about the same before and after the intervention. It therefore appeared unlikely that the ability level of the students had changed.

3. Compared to Change without the Intervention

There must also be some credible estimate of conditions that would be in effect without the intervention through the use of control groups, comparison groups or other appropriate standard such as a time series design.

The most severe barrier for adult education programs is that all members of the target groups are allowed to participate in the innovative program to be measured. If this condition is not compensated for, there is no basis on which to measure differences in levels of achievement. Statistical compensation for this and other conditions can be achieved through variations in the evaluation designs.

Sound conclusions rest on three steps:

a) Measuring the change in participants.

b) Measuring the change in absence of the programs.

c) Comparing the two changes.

VALID EVALUATION DESIGNS

Description of various statistically vaild evaluation designs, in descending order of credibility, and the conditions under which they would be used follow:

1. Random Selection Design

This design is used if individuals can be randomly selected and assigned to either the participant group or the non-participating control group. The random selection and assignment of individuals to either group assures the statistical equivalence of participants and non-participants.

2. Delayed Random Selection Design

There is no control group if everyone participates in the program. A comparison group is formed by randomly scheduling some potential participants to start the second cycle of a training program. The selected participants will begin instruction following completion of instruction for remaining participants. The outcomes of the group receiving the treatment first can be compared to the outcomes of the delayed group.

3. Varied Instruction Time Design

A second compromise to the basic control group design, when all members of the target group participate in the program, consists of scheduling individuals or groups of individuals to participate to varying degrees. If two or more groups receive different amounts of instruction, this is sufficient to define an instructional variable.

Random assignment of individuals to the various groups is essential to assume equivalence. The instructional variable can be related to individual change measures through the use of a statistical technique known as correlation or regression analysis.

4. Matched Groups Design

A third compromise to random assignment of individuals into treatment and control groups is to compare intact (pre-existing) groups who are similar in all relevant characteristics. The evaluator could seek to compare two different educational communities, one participating in the adult or vocational education program and one not participating in the program. The matching of groups should be conducted on an individual basis; that is, for each program participant a "twin" is matched from the comparison group.

A variation of this design is to compare two groups participating in different educational programs, (e.g. one traditional versus one innovative). In this case the incremental effectiveness is evaluated (i.e. the benefits accruing to participants over and above the benefits accruing from another program).

5. Intact Groups Design

This design involves the comparison of intact (pre-existing) groups but the comparison group may differ considerably from the treatment group in one or more relevant characteristics. Statistical adjustments must be

made with respect to the sources of non-equivalence between groups. Because of the complexity of making these adjustments they should be made by a qualified statistician.

6. Delayed Intact Groups Design

This design is the same as the intact groups design except that the comparison group eventually participates in the adult or vocational education program, after evaluation activities are completed. The comparison group is established by delaying the onset of the program for one of the groups.

7. Varied Instruction Time Design with Intact Groups

This design is used whenever Varied Instruction Time Design is appropriate but it is not possible to randomly assign individuals to treatment groups and as a result the groups are not truly equivalent. The estimate of the relationship between the instructional variable and the performance of individuals would be statistically adjusted for all sources of non-equivalence of the varying groups.

8. Selection Groups Design

This design is used when selection of members of the participant group and of the comparison group is made on the basis of a single educational criterion. Through a statistical technique known as "regression

discontinuity" the post-test performance of the groups may be compared. Two post-test scores are statistically projected to represent the post-test performance of two hypothetical individuals achieving the same selection score. However, one has participated in the program and the other has not. The difference between these two projected scores reflects the effectiveness of the program.

9. Norm-Referenced Design

This design is used when there is no comparison group and it is not possible under any circumstances to locate one. If standardized tests are used with nationally normed scores available, the pre- and post-test scores of program participants can be compared to the performance of a nationwide sample of individuals. It is especially important to provide documentation of the initial status or the expected growth rate of the participants in the absence of the intervention.

10. Time-Series Design

This design is used when a single program group is being evaluated in the absence of any comparison group, including a national norm group. An acceptable procedure is to examine the change of program participants over multiple points in time, before, during, and after the beginning of instruction.

When evaluation designs do not involve a comparison group, but the performance of the treatment group is compared with some norm or expectation it is especially important to provide documentation of the initial status or the expected growth rate of the participants in the absence of the interventions.

B. RELIABILITY

A reliable measure is consistent in its measurement, time after time. Few evaluators would unquestioningly accept the result of any single, small-scale study as adequate evidence of the success of any intervention regardless of the level of statistical significance. There should always be at least one replication or parallel study. If, for example, comparable results are observed in two or more classrooms, or in two or more successive years, or both, results become much more credible. This type of consistency of finding not only helps to establish statistical significance and intuitive credibility, it is also directly relevant to the transportability criterion.

To summarize, a measure which possess both validity and reliability as defined above is statistically significant.

In addition a proposed intervention must also establish educational significance.

II. EDUCATIONAL SIGNIFICANCE

Educational significance is not a matter of statistics but relies on judgment. The Joint Dissemination and Review Panel (J.D.R.P.) considers the following three factors in judging the educational significance of a program.

A. Size of the effect

B. Importance of the area of change

C. Reasonableness of cost

The J.D.R.P., in making a judgment, considers the first two factors together, assessing whether or not there is a reasonable balance between them. The chance that a small gain would be considered educationally significant is higher in a broad or educationally important area than in a narrow or less important area.

A. Size of the Effect

In weighing the size of educational effects one widely applied statistical rule of thumb is that the effect must equal or exceed some proportion of a standard deviation -- usually one-third, but at times as small as one-fourth -- to be considered educationally significant. Another statistical criterion is rate of growth that will produce a post-test percentile standing that exceeds the present percentile standing by one standard error of measurement.

B. Importance of the Area of Change

There is a parallel between the breadth of focus of educational interventions and that of the measures used to assess their impact. Standardized achievement test scores are the most widely known and accepted measures for use in education evaluations. They have, however, been justifiably criticized on several counts. From an evaluator's viewpoint, the most relevant criticism is that they do not measure what is being taught. On the other hand, most people concede that such instruments do measure the ability to read and do arithmetic.

The ability of any test to measure change will be directly related to how relevant the test items are to the content of the instruction.

But, it is not necessary to use achievement tests to establish the educational significance of an intervention. The most convincing evidence of success in a dropout-prevention program, for example, would simply be statistics showing a decrease in the number of students dropping out. Similarly, change in adult and vocational education programs might be measured in terms of job placements, starting salaries, rates of advancement on the job, etc. Obviously achievement tests could also be used.

C. Reasonableness of Cost

Another factor entering into the consideration of educational significance is the matter of cost. Because resources are limited, more people can be served by low-cost interventions than by high-cost interventions. If they are available, cost-benefit figures should be presented. Cost-benefit can be defined as the amount of money which will be realized (i.e. received or saved), over a specific period of time, because of the operation of the program, in relation to the amount of money spent to operate the program.

To summarize, an educationally significant effect is one of nontrival magnitude, in a content area generally accepted as important, which can be achieved at a reasonable cost.

III. REPLICABILITY

Statistical significance may reassure us that project results were no fluke, but that still does not guarantee that the intervention will be effective when replicated in other settings. In order to determine the likelihood that the same products or practices, when used elsewhere, will produce results similar to those obtained at the original site, the panel considers the following four factors:

A. Uniqueness of Project Setting

The project setting should not be so unique that the project could not be replicated elsewhere. An intervention that works in an environment seldom found elsewhere may be deriving its effectiveness solely from

that environment and without further evidence of replicability would not be a good candidate for J.D.R.P. approval.

B. Project Staff

Although the J.D.R.P. is concerned about the likelihood that an educational intervention will work equally well in another setting, few evaluations are designed to prove this. The primary focus of the evaluation is, as it should be, upon the effectiveness of the intervention as it was carried out. But, whenever it appears likely that one or more rare individuals exerted an influence that typical school personnel could not duplicate, the replicability of the intervention is in question. There are various indicators of generalizability that can be provided without going to the trouble and expense of conducting a replication study. One technique is to involve more than one class and teacher in the original project wherever possible. Also, it is more convincing to select teachers randomly to implement a new approach that to use those who volunteer. The need is to provide evidence that teachers who carried out the intervention were not unusual, so that one could expect teachers elsewhere to get similar results if they use the same products or procedures. If the project involves other staff members—administrators, project directors, or specialists — the same procedures should be employed.

C. Participants

Similar considerations apply with regard to participants. The more there are, the better. Choosing them at random provides a more convincing case for replicability than does using volunteers. If this is not possible, it is a good idea to collect any available evidence that will support claims that those who participated were not different from potential participants anywhere else, and that their performance was a typical result of the intervention, rather than a unique response to it by unusual participants.

D. Replicability of Essential Components

Some evidence must be presented that essential components have been identified and that these can be replicated elsewhere. Some examples of these components might be teacher training, parental involvement, individualization of instruction, commercially available curricula.

To summarize, setting, staff, participants, and essential components should not be so unique that they could not be replicated elsewhere

V. BELIEVABILITY AND INTERPRETABILITY

A. Consistency of Factual Data in Narrative and Tables

One of the most telling signs of a flawed evaluation is the presence of inconsistencies in the data. An obvious problem is lack of agreement between numbers reported in the text and the tables, or among tables.

Another is inconsistencies in the calculations.

Lack of agreement between numbers could be the result of a typographical error. It could be an attempt to gloss over disappointing data. If fewer pupils were tested than were served, it could be the result of planned sampling that was part of the evaluation design and attrition may have left a biased sample at post-test time. Errors in calculations may simply be mistakes, or there may be an attempt to make a "right" answer out of wrong data. Any errors, however, tend to detract from the overall credibility of the submission.

B. Completeness of Data

Lack of complete data also precludes accurate interpretation of an evaluation. Sometimes submissions omit important information such as the names, form, and levels of tests used; the testing times, the number of students tested; or the number of students served.

Data may be presented on only some of the measures that were administered. Failure to report all of the data can arouse the suspicion that those not reported were somehow unfavorable. Whatever the reason, if information is missing, the evidence cannot be properly interpreted or taken at its face value.

In addition to basic information about the data that were collected, there should be a complete and accurate presentation of the analysis of these data. Types of scores should be clearly identified, e.g. raw scores, publishers' standard scores, Normal Curve Equivalents (NCEs), etc. Summary statistics should include both "means" and "standard deviations." Each time scores are reported, sample sizes should also be reported. When statistical tests are used, they should be clearly identified; the rationale for their use, if not obvious, should be presented; and any assumptions made should be explained.

A major defect in some evaluation designs, particularly norm-referenced designs, is the "regression effect" error. For example, if participants were selected for a remedial reading project on the basis of their low scores on the XYZ Reading Test, and if those same scores were then used to figure the average pre-project status of the students, the gains attributable to the intervention would be overestimated. Unless the evaluation report clearly states that different tests were used for selection and for pretesting, the reviewer has no assurance that this error was avoided. It is important to specify clearly that the scores used for selection of participants were not the same as those used for measurement of pre-intervention status. They must not be the same. (Scores from an alternate form of the same test, however, are perfectly acceptable.)

C. Objectivity Maintained in Gathering Data

Another important point to stress in the presentation of evaluation results is the objectivity of the data. Wherever it is possible for data to be contaminated, the write-up should describe measures taken to make sure that this did not occur. For example a common source of problems is the procedure followed in testing. When tests are administered by persons with a stake in the outcome -- such as staff members of a project, or those with a close personal relationship to the subjects, those test results are suspect. The belief is that the test administrators could have influenced the student's performance in some intangible and perhaps totally unintentional way. They may have given extra directions, allowed more time, or deviated in some other way from the instruction for test administration. If the test required judgments or ratings by the administrator, their objectivity would be seriously in doubt.

To make it clear that there were no irregularities in testing, an evaluation should specify who gave the tests, and under what conditions. For example, "Each participating class was given the XYZ Reading Test on May 1 (the same date that the national norm group was tested). The test was administered by teachers, who had been thoroughly trained in the publisher's instructions for proper use of the tests. Although it was not possible to obtain outside test administrators, the teachers were rotated so that they tested each other's classes, not their own."

To summarize, in order to meet the tests of believability and interpretability, the data pertaining to a project must be consistent, complete, and objective.

The standards of statistical significance, educational significance, replicability and believability are used by the Joint Dissemination and Review Panel (J.D.R.P.) to judge the effectiveness of projects. Projects meeting these standards will be recognized by the U.S. Department of Education as worthy of adoption.

EVALUATION DESIGN CHECKLIST

OBJECTIVES	1. What objectives are planned as a result of the intervention?
PROCEDURES	2. Is the project methodology adequately described? 3. Will the procedures be consistently employed by all staff?
EVALUATION DESIGN	4. What kind of change is intended? a. educational? b. attitudinal? c. behavorial? 5. Will the change be logically measured by relevant pre-testing and post-testing? 6. Will precautions be taken in the design of the evaluation to neutralize outside influences such as side effects from the experiment or maturation?
MEASURES	7. Will the pre-test instrument be different from the instrument used for selection? 8. How will the estimate of conditions without the intervention be measured ? a. control group? b. comparison group? c. other standard? 9. Will the comparison group be reasonably like the treatment group or will it be matched statistically?
CONDITIONS	10. Will the testing conditions be the same for both groups?
STATISTICS	11. Will the statistical analyses be appropriate to the evaluation design?
RELIABILITY	12. Is more than one observation planned to strengthen the case for reliability?

13. How will chance be ruled out as a possible cause of the change?

EDUCATIONAL SIGNIFICANCE

14. Is the change expected to be educationally significant as related to:
    a. Size of the effect?
    b. Importance of the area of change?
    c. Resonableness of cost?

REPLICABILITY

15. Will it be possible to replicate the project in another location?
    a. Is the setting neutral in effect?
    b. Can staff substitutes be found elsewhere?
    c. Are the participants typical enough to keep the project unaffected?
    d. Can essential components such as curriculum or teacher training courses be easily replicated?

BELIEVABILITY

16. Will the evidence presented be believable and interpretable?
    a. Will the statistics in text and tables be consistent with each other?
    b. Will the data be complete?
    c. Will objectivity be preserved in gathering the data?

REFERENCES

Tallmadge, G.k. The Joint Dissemination Review Panel: Ideabook. Washington, D.C.: National Institute of Education/DHEW, September 1977.

U.S. Department of Education. Educational Programs That Work. San Francisco, CA: Far West Laboratory, Fall 1978.

U.S. Office of Education. Developing Effective Evaluations for Adult Education Programs: A Handbook for Administrators and Evaluators. Harrison, New York, MAGI Educational Services, Inc. November, 1979.

Hamilton, Jack A. "Promoting National Dissemination of Exemplary Adult Basic Education Projects." Lifelong Learning: The Adults Years March, 1981.